

Språkteknologi (SV2122)

Föreläsning 9: Översättning



Richard Johansson

`richard.johansson@svenska.gu.se`

26 februari 2014

översikt

- ▶ inledning: vad är översättning?
- ▶ djupa metoder:
 - ▶ datorn läser och “förstår” texten
 - ▶ generera därefter en översättning utifrån “betydelsen”
- ▶ ytliga metoder:
 - ▶ observera hur ett ord eller en fras brukar översättas i korpusar
 - ▶ generera översättningen genom att återanvända fraser
- ▶ utvärdering av översatt text
- ▶ något om verktyg

vad är översättning?

- ▶ givet en text skriven på ett språk (**källspråket**)
- ▶ ...producera en text på ett annat språk (**målspråk**)
- ▶ ...så att de har samma **betydelse**



behov av översättning

- ▶ vardagligt, t.ex. jag vill förstå en websida på ryska
- ▶ översättning av t.ex. skönlitteratur: förmodligen en liten del av den faktiska översättning som görs
- ▶ exempel på områden där översättning massproduceras:
 - ▶ samhällsinformation t.ex. på Sveriges 5 minoritetsspråk
 - ▶ flerspråkiga politiska organisationer t.ex. EU, Finland
 - ▶ manualer
 - ▶ lagar, patent
 - ▶ underrättelseverksamhet



Introduction

- ▶ Automatic translation: a computer program that translates a text in one language (**the source**) into another language (**the target**).
- ▶ This is one of the most high-profile areas of NLP, and perhaps its most classical problem
- ▶ Challenges:
 - ▶ How to preserve the meaning? [**adequacy**]
 - ▶ How to respect the grammar of the target language [**fluency**]

hjälpmedel för översättare

- ▶ även den som vill göra sin översättning manuellt kan ha nytta av en del olika verktyg:
 - ▶ översättningsminnen
 - ▶ se t.ex. <http://sv.glosbe.com/sv/en/konstig>
 - ▶ konkordanser
 - ▶ terminologidatabaser
 - ▶ ...



översikt

“djupa” metoder

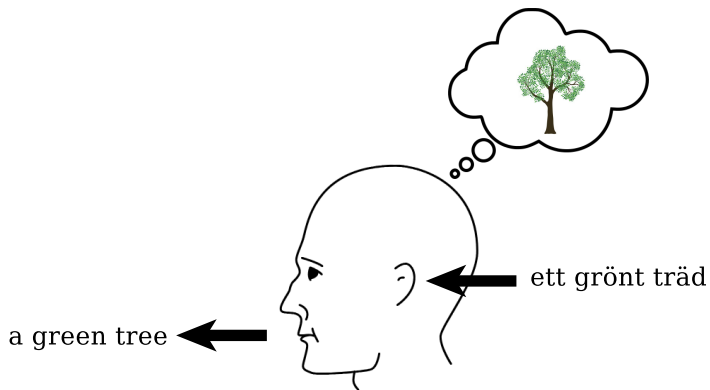
“ytliga” metoder

utvärdering

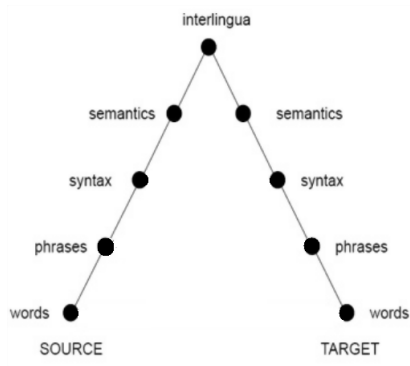
fortsättningen



idealiserad bild av översättningprocessen



klassisk maskinöversättning



ett konstruerat exempel

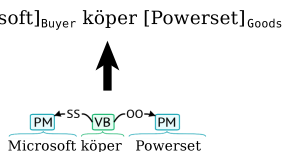
Microsoft köper Powerset



ett konstruerat exempel



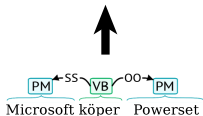
ett konstruerat exempel



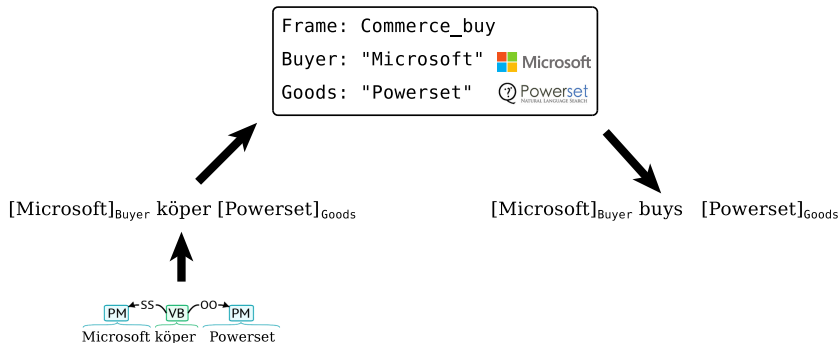
ett konstruerat exempel

Frame: Commerce_buy
Buyer: "Microsoft"  Microsoft
Goods: "Powerset"  Powerset
INTEGRAL LANGUAGE SEARCH

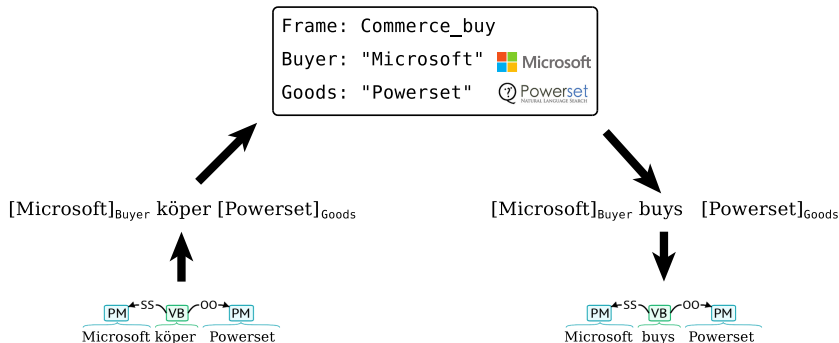
[Microsoft]_{Buyer} köper [Powerset]_{Goods}



ett konstruerat exempel



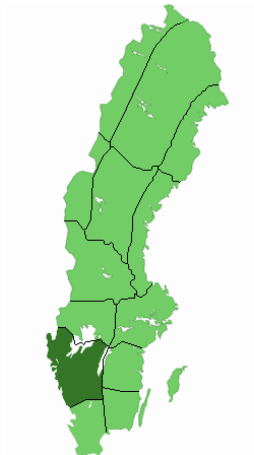
ett konstruerat exempel



mellanrepresentationer i praktiken

- ▶ i praktiken är det svårt att göra någon allmänt användbar mellanrepresentation
 - ▶ ...och att analysera betydelse i obegränsad text är ett öppet forskningsproblem!
- ▶ däremot kan det fungera bra om vi vet att texterna kommer från något begränsat område

Prognos för nordvästra Götaland



2014-02-24 KL 12:08

Utsikter till tisdag kväll

En hel del moln men främst i söder kan solen tränga igenom täcket lite grand. I natt lokalt ca 0. I morgon minskar dock molnigheten och det blir en ganska solig dag med 4-9 grader. Sydlig vind, frisk vid kusten.

Meteorolog: Lars Knutsson

exempel: projektet MOLTO

- ▶ `http://www.molto-project.eu`
- ▶ `http://www.molto-project.eu/cloud/gf-application-grammars`

översikt

“djupa” metoder

“ytliga” metoder

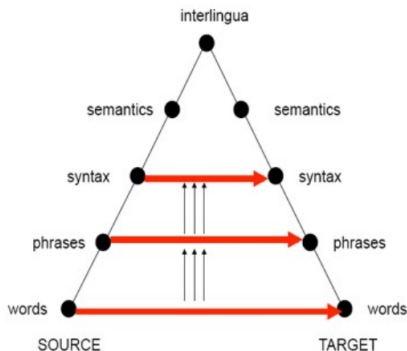
utvärdering

fortsättningen



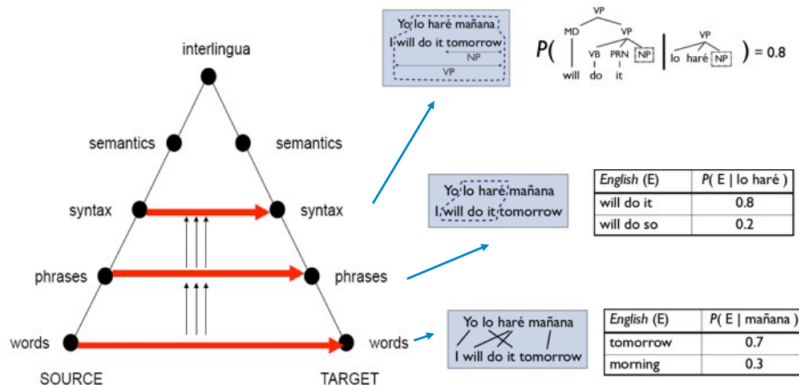
korpusbaserad maskinöversättning

- Statistical, corpus-based MT: observe translated corpora, translate new text by recombining fragments of previous examples



korpusbaserad maskinöversättning

- Statistical, corpus-based MT: observe translated corpora, translate new text by recombining fragments of previous examples



varför korpusbaserad maskinöversättning?

- ▶ praktiskt: det har visat sig oerhört svårt att beskriva “betydelse” på ett praktiskt sätt förutom i begränsade områden t.ex. väderleksrapporter
- ▶ lingvistiskt:
 - ▶ språkproduktion består till stor del av återanvändning av halvfabrikat (jfr t.ex. konstruktionsgrammatik)
 - ▶ det finns mycket mer godtycke än vad vi kanske tänker oss!
- ▶ idén om återanvändning av fragment är besläktad den om översättningsminnen

Statistical Machine Translation

- ▶ One of the first research areas in LT; since late 1940s.
- ▶ Weaver (1949): *When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."*
 - ▶ Suggested a “noisy channel” model inspired by cryptography, very similar to modern SMT
 - ▶ we can think of translation as a decryption problem; we are given a signal and want to find the underlying message
 - ▶ Not feasible in the early days of computing, but took off in the 1980s-1990s with the availability of fast computers and availability of data
 - ▶ <http://www.mt-archive.info/Weaver-1949.pdf>

Parallel Corpora

- ▶ Statistical models are estimated from **parallel corpora**
 - ▶ A corpus of “the same” text in two languages
 - ▶ Normally hand-aligned on the sentence level
- ▶ IBM research team: Canadian Hansards, English–French
- ▶ Europarl, <http://www.statmt.org/europarl>

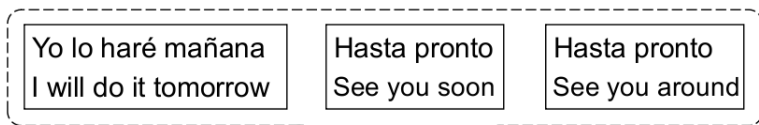
exempel: Europarl

- ▶ `http://spraakbanken.gu.se/korp/?mode=parallel`

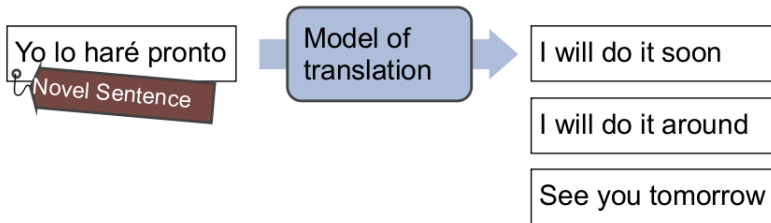


Corpus-based machine translation

Sentence-aligned parallel corpus:



Machine translation system:



IBM models

- ▶ IBM models are “simple” word-based translation models
- ▶ There are several variants (IBM models 1–6)
- ▶ Introduces a crucial concept: **word alignment**
- ▶ Word-based translation is usually too simple to be useful
 - ▶ May still be useful in very closely related language pairs such as Swedish–Danish.
- ▶ However, these models are still important today: word alignments and word translation probabilities are used for computing
 - ▶ Phrase tables and phrase translation probabilities in phrase-based translation
 - ▶ Grammar rule translation probabilities in syntax-based translation

matematiken i IBM-metoderna

- ▶ en central idé i IBM-metoderna och många andra statistiska översättningmetoder är att man använder två sorters sannolikheter:
 - ▶ **översättning av ord**: t.ex. vad är sannolikheten att *vård* översätts till *care*?
 - ▶ **ordföljd**: t.ex. vilken ordföljd är sannolikast, *nödvändigt vård* eller *nödvändig vård*?
- ▶ man brukar formulera översättningen med följande ekvation:

$$\text{översättning} = \arg \max_T P(T)P(S | T)$$

- ▶ ...där $P(S|T)$ betyder översättningssannolikhet, $P(T)$ ordföljdssannolikhet, och $\arg \max_T$ "hitta den översatta mening som maximerar sannolikheten"

Word alignment example

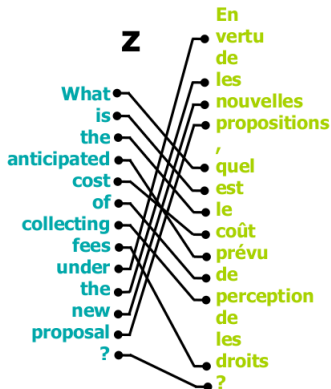
X

What is the anticipated
cost of collecting fees
under the new proposal?

En vertu des nouvelles
propositions, quel est le
coût prévu de perception
des droits?

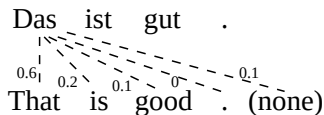


Z



hur går ordlänkningen till?

- ▶ en parallellkorpus är översatt mening för mening
- ▶ hur länkar vi ord till ord om vi inte vet vad orden betyder?
- ▶ idé:
 - ▶ titta vilka ord som brukar förekomma samtidigt i översatta meningar
 - ▶ ju oftare de samförekommer, desto säkrare blir vi på att de ska länkas



exempel på länkningar av enskilda ord i Europarl

- ▶ *vård*, totalt 982 förekomster:

- ▶ 385 *care*
- ▶ 79 *treatment*
- ▶ 51 ***health***
- ▶ 43 *healthcare*
- ▶ 17 ***the***
- ▶ 10 *caring*

statistik i ordbaserad översättning (IBM)

- ▶ We are given a set of sentence pairs in a parallel corpus.
- ▶ How do we estimate the probabilities the IBM model?
- ▶ Compute the word alignments ...
- ▶ ...then we can estimate word translation probabilities:

$$P(vård \rightarrow care) = \frac{\text{antal } vård \rightarrow care}{\text{antal } vård} = \frac{385}{982}$$

vård av släktingar
|
care of relatives

statistiken i ordbaserad översättning: ordföljder

- ▶ ordföljdssannolikheterna tas fram genom att räkna vilka ordföljder som är vanligast i korpusar
- ▶ den kan hjälpa oss att välja mellan olika tänkbara översättningar
 - ▶ t.ex. *nödvändig vård* förekommer 3 gånger i Europarl;
nödvändigt vård inte alls
- ▶ fördel med en separat ordföljdsdel: statistiken i den kan konstrueras med enspråkiga korpusar

example: ranking translations with a language model

Τη σερβίρουμε χλιαρή.

⇒

To serve warmly.
To be placed lukewarm.
Serve it warm.

...

exempel på hur översättningen går till

- ▶ exempel *a small town*

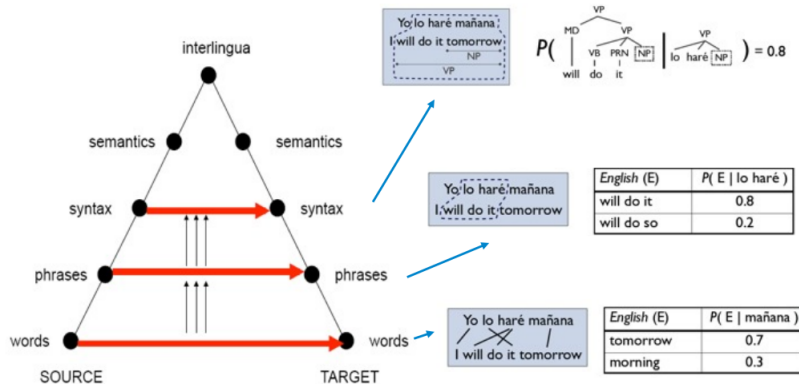
$a \rightarrow \{ en, ett, \dots \}$

$small \rightarrow \{ små, liten, litet, lilla, mindre, \dots \}$

$town \rightarrow \{ stad, staden, städer, samhälle, hemstad \dots \}$

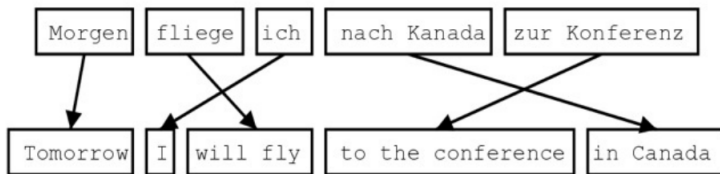
- ▶ välj den ordföljd som ger högst sannolikhet
 - ▶ översättningssannolikheter för översättning av enskilda ord
 - ▶ ordföljdsannolikheter för svenska ordföljder
- ▶ uppgiften att välja den översättning som ger högst sannolikhet brukar kallas **avkodning** (*decoding*)
- ▶ observera likheten med ordklassanalys i det här enkla fallet!

översikt igen



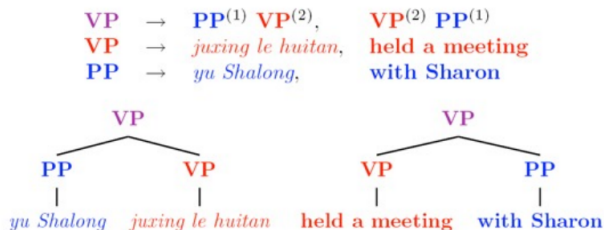
Phrase-Based Statistical Translation

- ▶ Phrase-based translation models are a generalization of the word-based models, where translation, distortion and fertility models are defined on phrases instead of words (see, e.g., Koehn, Och and Marcu 2003).
- ▶ Phrases need not be linguistically motivated but can be arbitrary n-grams in the two languages.



Syntax-Based Statistical Translation

- ▶ Synchronous context free grammars (SCFGs): context-free grammar in two dimensions
- ▶ Generates pairs of strings/trees simultaneously
- ▶ Translation: parse the source sentence, simultaneously generate the target sentence



Software for statistical MT

- ▶ GIZA++: Implements IBM models 1–5. Useful for computing word alignments (<http://code.google.com/p/giza-pp>)
- ▶ Moses: Reference implementation and research platform for standard phrase-based and syntax-based SMT algorithms (<http://www.statmt.org/moses>)
 - ▶ här finns också till ett par demonstrationssystem som byggts med hjälp av Moses – tyvärr verkar inte det finska systemet funka...
 - ▶ <http://www.statmt.org/moses/?n=Public.Demos>



översikt

“djupa” metoder

“ytliga” metoder

utvärdering

fortsättningen



Evaluation of machine translation systems

- ▶ MT evaluation is hard!
- ▶ Human judgments:
 - ▶ Fluency
 - ▶ Adequacy
- ▶ Examples:

Reference: Yesterday, stock and commodity prices fell on the world's markets.

Output 1: Stock markets and commodity markets fell globally yesterday.

Output 2: The stock market fell in Zurich.

Output 3: Around globe stock market, and and also, commodities fall yesterday.

Output 4: Market and win ball rolling yesterday around electronic highly.

Automatic evaluation methods

- ▶ Human judgments take too much time
- ▶ Automatic judgment: BLEU
 - ▶ P1: unigram precision
 - ▶ P2, P3, P4: bi-, tri-, 4-gram
 - ▶ Weighted mean
 - ▶ Brevity penalty
- ▶ BLEU is controversial! People often complain but no alternative is widely accepted.

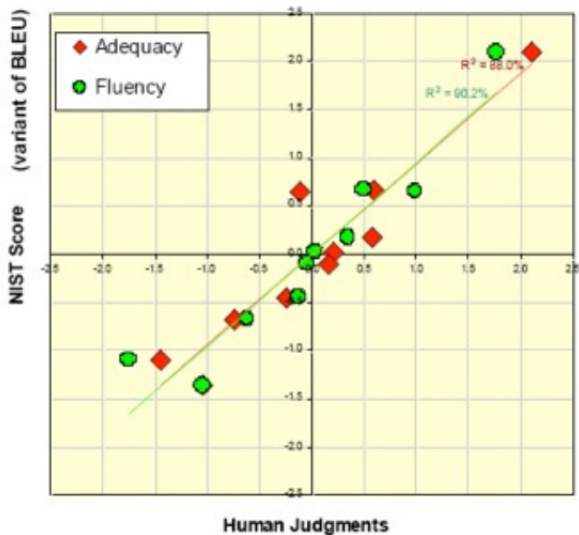
Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Machine translation:

The American [?] international airport and its the office an receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

Do automatic evaluation methods make sense?



utmaningar, t.ex. pronomenerferens

- ▶ det finns en hel del lingvistiska problem vid översättning som vi inte vet hur vi ska lösa, oavsett om vi använder “djupa” eller “ytliga” metoder
- ▶ ...det verkar som om det i sådana fall krävs att vi faktiskt förstår texten!

The house was red. It was Peter who built it.

exempel på utmaningar: ordbetydelse

La borsa di studio è cumulabile con il sussidio (Provvidenza) o con altra borsa erogata da altro ente?

The scholarship will be combined with the subsidy (Providence) or other bag delivered by another institution?

Stipendiet kommer att kombineras med bidraget (Providence) eller annan påse som levereras av en annan institution?



översikt

“djupa” metoder

“ytliga” metoder

utvärdering

fortsättningen



datorövning 2

- ▶ använda verktyg för ordklassanalys och syntaktisk analys och förstå varför de beter sig som de gör
- ▶ fredag i T225



nästa föreläsning

- ▶ datorstödd språkinlärning
- ▶ onsdag, **C452** (Humanisten)
- ▶ Elena föreslår att den som har en egen laptop tar med sig den

